Comment on "Invidious Comparisons: Ranking and Selection as Compound Decisions"[*]

Magne Mogstad
Department of Economics
University of Chicago
Statistics Norway & NBER
magne.mogstad@gmail.com

Joseph P. Romano
Departments of Statistics and Economics
Stanford University
romano@stanford.edu

Azeem M. Shaikh
Department of Economics
University of Chicago
amshaikh@uchicago.edu

Daniel Wilhelm
Department of Economics
University College London
d.wilhelm@ucl.ac.uk

January 19, 2022

[*] The second author acknowledges support from the National Science Foundation (MMS-1949845). The third author acknowledges support from the National Science Foundation (SES-1530661). The fourth author acknowledges support from the European Research Council (Starting Grant No. 852332). We thank Jiaying Gu and Roger Koenker for applying their procedures to the dataset in Section 5. We also thank Yuehao Bai, Max Tabord-Meehan and Andres Santos for helpful comments.

1 Introduction

In their paper, "Invidious Comparisons: Ranking and Selection as Compound Decisions," the authors, Gu and Koenker, masterfully develop an empirical Bayes approach to the problem of choosing the "best" (or, equivalently, "worst") of n populations indexed by $i = 1, \ldots, n$, where "best" is measured in terms of a feature of interest that is denoted by θ_i. There is, of course, abundant motivation for considering this problem in the economics literature. Two especially compelling examples are provided by recent interest in ranking teachers according to measures of teacher value-added (Chetty et al., 2014a,b) and ranking neighborhoods according to measures of intergenerational mobility (Chetty et al., 2014c; Chetty and Hendren, 2018a,b).

In this comment on their paper, we first briefly summarize in Section 2 the most general framework for the problem that they consider. We then specialize their framework so as to highlight a connection (already noted by the authors themselves) to the multiple testing literature. By doing so, we facilitate a comparison in Section 3 with a closely related multiple testing problem considered in Mogstad et al. (2020). Throughout our discussion, in addition to some key distinctions in the asymptotic framework employed in each paper, we emphasize two additional differences: differences in the choice of error rate for each multiple testing problem and differences in the way in which the θ_i are treated. In Section 4, we discuss the practical importance of these differences. To further emphasize what we feel are important distinctions in the interpretation of the results provided by the two methods, we cast this discussion in the context of the aforementioned teacher value-added example. In short, we argue that it may be politically more palatable to make consequential decisions, such as deciding which teachers lose their jobs, when employing a more stringent error rate that holds for each value of $\theta = (\theta_1, \ldots, \theta_n)$ (rather than a more permissive error rate that only holds in expectation over θ). By doing so, it is possible to defend these decisions with statements such as "with high probability, *no* teachers were fired incorrectly." Finally, in Section 5, we conclude with a comparison of applications of their methodology and ours to data from Chetty et al. (2018). Motivated by the "Creating Moves To Opportuntiy" experiment described in Bergman et al. (2020), we consider selecting the "best" commuting zones in terms of a measure of intergenerational mobility, to which families may relocate.

2 Overview of Gu and Koenker (2021)

In the setting of Gu and Koenker (2021), the features of interest θ_i are, of course, unknown, but it is assumed that for each population repeated measurements of the underlying θ_i are observed. More precisely, the observed measurements are given by $Y_{i,t} = \theta_i + \sigma_i \epsilon_{i,t}$ with $\epsilon_{i,t}, i = 1, \ldots, n, t = 1, \ldots, T_i$ i.i.d. $\sim N(0, 1)$. A key assumption underlying the empirical Bayes approach is that these signals are further related in some way. For example, for part of their analysis, Gu and Koenker assume that $\theta_i, i = 1, \ldots, n$ are i.i.d. $\sim G$ and σ_i^2 is known, but later also consider the case where σ_i^2 is unknown and also modeled as being i.i.d. jointly with θ_i. A population i is formally defined to be among the "best" if $\theta_i \geq \theta_\alpha$, where θ_α is the $1-\alpha$ quantile of the marginal distribution of θ_i according to G. Gu and Koenker use the following loss function to discipline their decisions about which populations are in fact the "best":

$$L(\delta, \theta) = \sum_{1 \leq i \leq n} I\{\theta_i \geq \theta_\alpha\}(1 - \delta_i) + \tau_1 \left(\sum_{1 \leq i \leq n} \{I\{\theta_i < \theta_\alpha\}\delta_i - \gamma\delta_i\} \right) + \tau_2 \left(\sum_{1 \leq i \leq n} \delta_i - \alpha n \right) ,$$

where $\theta = (\theta_1, \ldots, \theta_n)$ and $\delta = (\delta_1, \ldots, \delta_n)$ indexes decisions about each population and $\delta_i = 1$ if and only if the ith population is selected as being among the "best." Here, α, τ_1, τ_2 and γ are parameters of the loss function to be specified by the researcher. The first term in the loss function disciplines false non-discoveries, meaning failing to select a population when it is in fact among the "best"; the second term disciplines false discoveries, meaning selecting a unit when it is in fact not among the "best"; and the third term disciplines the overall number of selected populations. The authors then choose δ so as to minimize the risk given by $E[L(\delta, \theta)]$, where, importantly, the expectation is over both θ_i and σ_i^2 as well as the measurements themselves. The decision rule obtained in this way is infeasible because it depends on G, which is unknown, but the authors assume that it is suitably estimable in their asymptotic framework, in which n tends to infinity. By replacing G with this estimate, a feasible decision rule is obtained, and, under certain assumptions, it performs approximately as well as the infeasible rule when n is large.

For the purposes of our discussion, it is convenient to assume $\tau_2 = 0$. As noted by Gu and Koenker (2021), when this is the case, for a suitably chosen value of τ_1, the problem is closely related to testing the family of null hypotheses

$$H_i : \theta_i \geq \theta_\alpha \text{ versus } K_i : \theta_i < \theta_\alpha \tag{1}$$

in a suitable way. In particular, for an appropriately chosen value of τ_1, the problem is equivalent to testing this family of null hypotheses so as to minimize false non-discoveries as measured by the quantity $E[\sum_{1 \leq i \leq n} I\{\theta_i \geq \theta_\alpha\}(1 - \delta_i)]$ subject to controlling the marginal false discovery rate, meaning $mFDR \leq \gamma$, where

$$mFDR = \frac{\sum_{1 \leq i \leq n} E[I\{\theta_i < \theta_\alpha\}\delta_i]}{\sum_{1 \leq i \leq n} E[\delta_i]} .$$

As shown by Genovese and Wasserman (2002), in the asymptotic framework considered by Gu and Koenker, in which n tends to infinity, the marginal false discovery rate is approximately equal to the more familiar false discovery rate, denoted by FDR and defined to be

$$FDR = E\left[\frac{\sum_{1 \leq i \leq n} I\{\theta_i < \theta_\alpha\}\delta_i}{\sum_{1 \leq i \leq n} \delta_i}\right] . \tag{2}$$

For later comparison, we reiterate one of our prior points: the expectations above are over both θ_i and σ_i^2 as well as the measurements themselves.

3 Connection with Mogstad et al. (2020)

We now describe a way in which the above problem connects to one pursued in Mogstad et al. (2020). In order to do so, it is helpful to highlight first some important differences in the frameworks employed by both papers. Mogstad et al. (2020) treat the θ_i as fixed, unknown parameters, whereas Gu and Koenker treat them as being random according to (the marginal distribution of) G; furthermore, Mogstad et al. (2020) employ an asymptotic framework in which n remains fixed, whereas Gu and Koenker employ one in which n tends to infinity in order to facilitate estimation of G. We note, however, that Mogstad et al. (2020) rely upon suitably well behaved estimators of θ_i and σ_i^2. For this purpose, the authors typically rely upon an asymptotic framework in which $\min_{1 \leq i \leq n} T_i$ tends to infinity, whereas Gu and Koenker allow this to remain fixed and instead require a normality assumption. As an estimator of θ_i, it is natural to employ

$\bar{Y}_i = \frac{1}{T_i}\sum_{1\leq t\leq T_i} Y_{i,t}$, which, under the assumptions of Gu and Koenker is distributed as $N(\theta_i, \sigma_i^2/T_i)$. To aid our exposition below, it is convenient to assume that σ_i^2 is known, but we emphasize that this is not required in general.

With these distinctions in mind, we note that the counterpart to the null hypothesis H_i in (1) is $\widetilde{H}_i : r_i/n \leq \alpha$, where

$$r_i = 1 + \sum_{1\leq j\leq n} I\{\theta_i > \theta_j\} \tag{3}$$

is the "rank" of the ith population. Mogstad et al. (2020) develop methods for testing the family of null hypotheses

$$\widetilde{H}_i : r_i/n \leq \alpha \text{ versus } \widetilde{K}_i : r_i/n > \alpha \tag{4}$$

in a way that controls the familywise error rate, meaning $FWER_\theta \leq \gamma$ for all values of θ, where

$$FWER_\theta = P\left\{\sum_{1\leq i\leq n} \delta_i I\{r_i \leq \alpha\} > 1\right\} . \tag{5}$$

The subscripting by θ here is intended to emphasize that the probability in (5) treats θ as fixed. We also note that the requirement that this holds for all values of θ is not innocuous: not only does the distribution of the measurements change with θ, but so too does the set of null hypotheses that are true (or false), i.e., the identities of the populations i for which $r_i/n \leq \alpha$. As explained, e.g., in Hall and Miller (2009), the possibility of (near) ties among the θ_i raises particular challenges. While it is not presented in this way in Mogstad et al. (2020), such a testing procedure can be obtained by constructing what the authors refer to as a confidence set for the p-best of level $1-\gamma$ with $p = \lfloor \alpha n \rfloor$ and rejecting $\widetilde{H}_i$ if and only if i is not contained in the resulting confidence set. Such a confidence set contains the identities of all populations i with $r_i/n \leq \alpha$ with probability at least $1-\gamma$. It is important to draw attention to the fact that the probability calculation in the definition of the familywise error rate involves uncertainty about only the measurements, not about the θ_i, which are treated as fixed and unknown. We also note that in practice (e.g., when σ_i^2 is unknown) control of the familywise error rate is only achieved asymptotically, but, under the assumptions maintained here, it may be controlled in finite samples. See also Bazylik et al. (2021) for other instances in which finite-sample validity may be achieved. Of course, it is possible to demand control of other error rates, such as the false discovery rate, but a similar caveat would apply: the expectation involved in the calculation of the false discovery rate would now only involve uncertainty about the measurements. Formally, one would require $FDR_\theta \leq \gamma$ for all values of θ, where

$$FDR_\theta = E\left[\frac{\sum_{1\leq i\leq n} I\{r_i \leq \alpha\}\delta_i}{\sum_{1\leq i\leq n}\delta_i}\right] . \tag{6}$$

As before, the subscripting by θ is intended to emphasize that the expectation in (6) treats θ as fixed. Methods for control of the false discovery rate remains an active area of research and is especially challenging when one seeks to incorporate information about dependence across the statistics by which each null hypothesis is being assessed. Such methods are especially salient here because r_i defined in (3) depends not only on θ_i, but also θ_j for $j \neq i$. Some relevant methodology is developed in Romano et al. (2008a).

4 Discussion

We now turn our attention to the practical importance of two distinctions highlighted by the above discussion – differences in the choice of which error rate to control as well as differences in the way in which the θ_i are treated. To facilitate this discussion, it is helpful to consider a specific example, such as selecting the "best" teachers in terms of teacher value-added. Here, θ_i is the value-added of the ith teacher and "best" means being among the top α fraction of all teachers in terms of θ_i. Teachers who are deemed as not being among the "best" may face consequences, including possibly losing their jobs (see, e.g., Hanushek (2011)).

Differences in the error rate: If one were to demand $FWER_\theta \leq \gamma$ for all θ as in Mogstad et al. (2020), then one obtains decisions such that the probability of incorrectly classifying any teacher as not being among the "best" is no more than γ. Informally, with probability at least $1 - \alpha$, *no* teachers are fired wrongly, regardless of the value of θ. Can an analogous statement be made if one were to demand $FDR_\theta \leq \gamma$ for all θ, where FDR_θ is defined in (6)? It is difficult to do so. To appreciate why, note that FDR_θ is the expected value of the fraction of total discoveries that are false – also known as the false discovery proportion, denoted by FDP_θ and defined to be

$$FDP_\theta = \frac{\sum_{1 \leq i \leq n} I\{r_i \leq \alpha\}\delta_i}{\sum_{1 \leq i \leq n} \delta_i} \ . \tag{7}$$

Unfortunately, control of the expected value of the FDP_θ does little to discipline its distribution. Some restrictions can, however, be obtained. Markov's inequality implies, e.g., that for any $0 < c < 1$,

$$P\left\{FDP_\theta > c\right\} \leq \frac{FDR_\theta}{c} \leq \frac{\gamma}{c} \ . \tag{8}$$

Hence, the probability of incorrectly classifying more than a fraction c of teachers as not being among the "best" is no more than γ/c. Informally, with probability at least $1 - \gamma/c$, the fraction of teachers who are fired wrongly is at most c, regardless of the value of θ. For this statement to be meaningful, however, we require that c is both small and large (relative to γ), which limits its practical importance. In our view, it may therefore be politically more palatable to make decisions about which teachers possibly lose their jobs when employing the familywise error rather than the false discovery rate. We note, however, that it is possible to target the left-hand side of (8) directly, i.e., to demand control of the tail probability of the false discovery proportion. By doing so, one may circumvent the shortcomings of the false discovery rate suggested by the discussion above. For a description of some such methods, see Lehmann and Romano (2005), Romano and Shaikh (2006a,b), Romano and Wolf (2007), Romano et al. (2008b) and Guo et al. (2014). For an empirical Bayes approach to the same problem, see Basu et al. (2021), who, by way of motivation for requiring control of the tail probability of the false discovery proportion, further illustrate in a simulation study that false discovery rate-controlling procedures may still permit a large fraction of total discoveries to be false quite frequently (see, in particular, their Figure 1).

Differences in the treatment of θ_i: Recall that the decisions obtained by Gu and Koenker ensure (at least approximately for large n) that $FDR \leq \gamma$, where FDR is defined in (2). As mentioned previously, an important aspect of this requirement is that the expectation in (2) is also over uncertainty in θ_i. Using the law of iterated expectations, it is straightforward to see that such a restriction may impose little discipline on $E[FDR|\theta]$, which corresponds more closely to the FDR_θ defined in (6). Needless to say, in light of the discussion above, this implies even less discipline on $P\{FDP > c|\theta\}$, where, by analogy with the definition in (7),

$$FDP = \frac{\sum_{1\leq i\leq n} I\{\theta_i \geq \theta_\alpha\}\delta_i}{\sum_{1\leq i\leq n}\delta_i} .$$

In our view, teachers may be more concerned with the behavior of these conditional quantities rather than their unconditional counterparts, especially when the stakes are high. After all, teachers may view their own θ_i as immutable and may view error rates that were only satisfied on average over possible values of θ_i as being irrelevant (or, indeed, even invidious) if their own job is at risk.

5 Empirical Illustration

In this section, we present a brief illustration of the above discussion in the context of ranking commuting zones (CZs) in the U.S. by a measure of intergenerational income mobility. The estimates of mobility and their standard errors are the same as the "correlational" estimates that Mogstad et al. (2020) select from the large dataset of Chetty et al. (2018). Suppose a family considers moving to one of the 100 most populous CZs. From these, the family wants to select a CZ whose mobility is among the ten highest. To this end, the family considers two approaches. First, it computes a confidence set for the 90-worst CZs at level $1-\gamma$ as in Mogstad et al. (2020).[1] Then the family selects all CZs that are outside of this confidence set. Second, it ranks CZs by their posterior tail probability as in Gu and Koenker (2021) using only the FDR constraint ($\tau_2 = 0$), $\alpha = 0.1$, and some value for γ.[2] Table 1 shows the CZs selected by the two procedures, denoted by "MRSW" and "GK", for different values of γ.

Since controlling the $FWER_\theta$ for all θ is more stringent than controlling the FDR, it is natural to expect GK to select more CZs. We find that MRSW selects only five CZs whereas GK selects more than the target of ten. Informally, the selection by MRSW guarantees that with probability approximately $1-\gamma$, *none* of the selected CZs (San Francisco, Salt Lake City, Honolulu, Boston and Minneapolis) can be among the worst 90 CZs (where this probability is calculated over different realizations of the data, but holding the mobility of each CZ fixed). Therefore, the family can be confident that a move to any of these CZs is indeed a move to a CZ with among the highest mobility. In contrast, GK selects a significantly larger number of CZs. Informally, their selection ensures that on average (where this average is over both different realizations of the data as well as different realizations of mobility for each CZ from a common distribution) no more than approximately γ fraction of the selected CZs are in fact among the worst. Therefore, the GK selection does not offer guarantees for the specific CZs selected to indeed be of high mobility. By analogy with our discussion in the preceding section, we believe families may view the MRSW selection as being more relevant for their decision about to which CZ to move, especially to the extent that mobility is a fixed feature for each CZ.

[1] The "DP" method described in Appendix F.2.

[2] As in Gu and Koenker (2021)'s empirical application, we consider the one-dimensional model in which $\theta_i \sim G$ and σ_i are treated as known. The procedure "KWs" from Gu and Koenker (2021, Section 7.1) with bandwidth 0.02 is used to estimate G.

		MRSW			GK		
rank	commuting zone	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.2$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.2$
1	San Francisco	×	×	×	×	×	×
2	Salt Lake City	×	×	×	×	×	×
3	Honolulu	×	×	×	×	×	×
4	Boston	×	×	×	×	×	×
5	Minneapolis	×	×	×	×	×	×
6	Toms River				×	×	×
7	Des Moines				×	×	×
8	San Jose				×	×	×
9	Scranton				×	×	×
10	Newark				×	×	×
11	Madison				×	×	×
12	Pittsburgh					×	×
13	New York						
14	Seattle						
15	Reading						
16	Manchester						×
17	Santa Barbara						×

Table 1: Selection of the "best" commuting zones in terms of intergenerational mobility among the 100 most populous commuting zones. MRSW selects commuting zones not in the level $1 - \gamma$ confidence set for the 90-worst; GK shows the commuting zones selected by the posterior tail probability with the $mFDR$ constrained at γ.

References

Basu, P., Fu, L., Saretto, A. and Sun, W. (2021). Empirical bayes control of the false discovery exceedance. *arXiv preprint arXiv:2111.03885.*

Bazylik, S., Mogstad, M., Romano, J. P., Shaikh, A. and Wilhelm, D. (2021). Finite-and large-sample inference for ranks using multinomial data with an application to ranking political parties. Working Paper 29519, NBER.

Bergman, P., Chetty, R., DeLuca, S., Hendren, N., Katz, L. F. and Palmer, C. (2020). Creating moves to opportunity: Experimental evidence on barriers to neighborhood choice. Working Paper 26164, NBER.

Chetty, R., Friedman, J. N., Hendren, N., Jones, M. R. and Porter, S. R. (2018). The opportunity atlas: Mapping the childhood roots of social mobility. Working Paper 25147, NBER.

Chetty, R., Friedman, J. N. and Rockoff, J. E. (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review*, **104** 2593–2632.

Chetty, R., Friedman, J. N. and Rockoff, J. E. (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review*, **104** 2633–79.

Chetty, R. and Hendren, N. (2018a). The impacts of neighborhoods on intergenerational mobility i: Childhood exposure effects. *The Quarterly Journal of Economics*, **133** 1107–1162.

Chetty, R. and Hendren, N. (2018b). The impacts of neighborhoods on intergenerational mobility ii: County-level estimates. *The Quarterly Journal of Economics*, **133** 1163–1228.

Chetty, R., Hendren, N., Kline, P. and Saez, E. (2014c). Where is the land of opportunity? the geography of intergenerational mobility in the united states. *The Quarterly Journal of Economics*, **129** 1553–1624.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64** 499–517.

Gu, J. and Koenker, R. (2021). Invidious comparisons: Ranking and selection as compound decisions. Tech. rep.

Guo, W., He, L. and Sarkar, S. K. (2014). Further results on controlling the false discovery proportion. *The Annals of Statistics*, **42** 1070–1101.

Hall, P. and Miller, H. (2009). Using the bootstrap to quantify the authority of an empirical ranking. *The Annals of Statistics*, **37** 3929–3959.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education review*, **30** 466–479.

Lehmann, E. and Romano, J. P. (2005). Generalizations of the familywise error rate. *The Annals of Statistics*, **33** 1138–1154.

Mogstad, M., Romano, J. P., Shaikh, A. and Wilhelm, D. (2020). Inference for ranks with applications to mobility across neighborhoods and academic achievement across countries. Working Paper 26883, NBER.

Romano, J. P. and Shaikh, A. M. (2006a). On stepdown control of the false discovery proportion. In *Optimality*. Institute of Mathematical Statistics, 33–50.

Romano, J. P. and Shaikh, A. M. (2006b). Stepup procedures for control of generalizations of the familywise error rate. *The Annals of Statistics*, **34** 1850–1873.

Romano, J. P., Shaikh, A. M. and Wolf, M. (2008a). Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test*, **17** 417–442.

Romano, J. P., Shaikh, A. M. and Wolf, M. (2008b). Formalized data snooping based on generalized error rates. *Econometric Theory*, **24** 404–447.

Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing. *The Annals of Statistics*, **35** 1378–1408.